Optimization and Large Scale Learning

Optimization lies at the heart of almost every machine learning algorithm. It is enjoying tremendous growth within our community, spurred largely by the specific needs of machine learning applications. Some key facets of these applications are: high-dimensional, noisy, and uncertain data; huge volumes of batch or streaming data; intractable models, low accuracy, reliance on distributed computation or stochastic approximations, and so on. Successfully handling these facets requires optimization techniques tailored to not only respect them but to aggressively exploit them.

These concerns motivate our overall research direction. It is worth noting that almost all contributions arose in response to the needs of one or more applications.

The gist of our agenda may be obtained by looking at the recent book [1] (MIT Press, 2011), or at the following workshops: (i) *"Optimization for Machine Learning"* (OPT2008–OPT2012; by S. Sra, S. Nowozin, and co-organizers); and (ii) *"Discrete optimization in machine learning"* (DISCML2009–2012, S. Jegelka et al.), which were organized by some of us as a part of the *Neural Information Processing Systems (NIPS)* conference. Both the book and the workshops have been successful platforms for interplay between the machine learning and optimization communities. Let us now list a few specific projects in line with our broad theme of optimization and machine learning; these projects are chosen to represent their highlighted subareas.

Large-scale regularized optimization Countless problems within machine learning and statistics can be formulated as regularized loss minimization. For typical convex 'loss-plus-regularizer' problems (e.g., lasso, group-lasso, trace norm lasso), we present a scalable method in [15]; while in [17], we present a new framework for large-scale nonsmooth, possibly nonconvex problems (e.g., dictionary learning, sparse matrix factorization, multiframe blind deconvolution). For linear and non-linear regularized kernel methods, characterized by a convex losses and quadratic regularizers, we derive two general classes of first order large scale methods, including coordinate descent techniques, and study their convergence properties in a unified framework [3]. A more specialized regularized problem is that with total-variation penalties, which we extend to complex multidimensional data in [8]. Going beyond first-order methods, we also study scalable Newton-type methods [2].

Core subroutines Complex optimization pipelines usually rely on fast core subroutines. In [4] we provide state-of-the-art algorithms for bound-constrained convex optimization; while in [5] we develop a high-performance method for nonnegative least squares. We present some well-tuned subroutines for nonsmooth group-sparse regularizers in [6].

Semidefinite optimization Semidefinite programming within machine learning is often used to solve relaxations of hard problems. We present a simple scalable technique in [16], though this has now been superseded by newer methods. We launch a new set of nonconvex yet globally optimizable positive definite problems in [7]; this has already benefited some of our applications in computer vision [11].

Fast search Indubitably, a key component of several practical learning systems is fast nearest neighbor search. Despite being an extremely well-researched area, in practice simpler approaches are still the most competitive [9, 10].

Discrete optimization Many application areas demand discrete models and solutions, and learning and inference become discrete optimization problems. In [1] we propose a novel method for analyzing the stability of linear programming relaxations for graph partitioning that arise in common clustering problems. More recently, we studied nonlinear combinatorial optimization problems with submodular cost functions, in particular scalable approximation algorithms [18, 18] that can be extended to an online setting [13]. Our models express a class of high-order graphical models with a particular structure [12], and, applied in computer vision, they significantly improve the state-of-the-art in segmenting difficult images [14].

Publications

Books

[1] S Sra, S Nowozin, and SJ Wright. *Optimization for Machine Learning*. Neural information processing series. MIT Press, Cambridge, MA, USA, 12 2011. 1

Book Chapters

[2] M Schmidt, D Kim, and S Sra. Projected Newton-type methods in machine learning. In S Sra, S Nowozin, and S J Wright, editors, *Optimization for Machine Learning*, pages 305–330. MIT Press, Cambridge, MA, USA, 12 2011. 1

Journal Articles

[3] F Dinuzzo. Analysis of fixed-point and coordinate descent algorithms for regularized kernel methods. *IEEE Transactions on Neural Networks*, 22(10):1576–1587, 10 2011. 1

[4] D Kim, S Sra, and IS Dhillon. Tackling box-constrained optimization via a new projected quasi-Newton approach. *SIAM Journal on Scientific Computing*, 32(6):3548–3563, 12 2010. 1

[5] D Kim, S Sra, and IS Dhillon. A non-monotonic method for large-scale non-negative least squares. *Optimization Methods and Software*, 2012. 1

[6] S Sra. Fast projection onto mixed-norm balls with applications. *Minining and Knowledge Discovery (DMKD)*, 2012. 1

[7] S Sra. The Stein divergence and metric on positive definite matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11 2012. 1

Articles in Conference Proceedings

[8] AJ Barbero and S Sra. Fast Newton-type methods for total-variation with applications. In L Getoor and T Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 313–320, Bellevue, Washington, USA, 2011. Omnipress. 1

[9] L Cayton. Efficient Bregman range search. In Y Bengio, D Schuurmans, J Lafferty, C Williams, and A Culotta, editors, *23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 243–251, Vancouver, BC, Canada, 2009. Curran. 1

[10] L Cayton. Accelerating nearest neighbor search on manycore systems. In *26th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2012)*, pages 1–10, Shanghai, China, 5 2012. 1

[11] A Cherian, S Sra, A Banerjee, and N Papanikolopoulos. Efficient similarity search for covariance matrices via the Jensen-Bregman logdet divergence. In DN Metaxas, L Quan, A Sanfeliu, and LJ Van Gool, editors, *13th International Conference on Computer Vision (ICCV 2011)*, pages 2399–2406, Barcelona, Spain, 2011. IEEE. 1

[12] S Jegelka and J Bilmes. Approximation bounds for inference using cooperative cut. In L Getoor and T Scheffer, editors, *28th International Conference on Machine Learning (ICML)*, pages 577–584, Bellevue, WA, USA, 7 2011. International Machine Learning Society. 1

[13] S Jegelka and J Bilmes. Online submodular minimization for combinatorial structures. In L Getoor and T Scheffer, editors, *28th International Conference on Machine Learning (ICML)*, pages 345–352, Bellevue, WA, USA, 7 2011. International Machine Learning Society. 1

[14] S Jegelka and J Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 1897–1904, Colorado Springs, CO, USA, 6 2011. IEEE. 1

[15] D Kim, S Sra, and I Dhillon. A scalable trust-region algorithm with application to mixed-norm regression. In J Fürnkranz and T Joachims, editors, *27th International Conference on Machine Learning (ICML)*, pages 519–526, Haifa, Israel, 6 2010. International Machine Learning Society. 1

[16] B Kulis, S Sra, and I Dhillon. Convex perturbations for scalable semidefinite programming. In D van Dyk and M Welling, editors, *Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 296–303, Clearwater Beach, FL, USA, 4 2009. MIT Press. 1

[17] S Sra. Scalable nonconvex inexact proximal splitting. In P Bartlett, FCN Pereira, CJC Burges, L Bottou, and KQ Weinberger, editors, *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 539–547, Lake Tahoe, Nevada, USA, 2012. 1

Theses

PhD Theses

[18] S Jegelka. *Combinatorial Problems with Submodular Coupling in Machine Learning and Computer Vision*. Phd, ETH Zürich, 2012. 1